Interactive digital slides with heat maps: a novel method to improve the reproducibility of Gleason grading

Egevad, L ; Algaba, F ; Berney, D M ; Boccon-Gibod, L ; Compérat, E ; Evans, A J ; Grobholz, R ; Kristiansen, G ; Langner, C ; Lockwood, G ; Lopez-Beltran, A ; Montironi, R ; Oliveira, P ; Schwenkglenks, M ; Vainer, B ; Varma, M ; Verger, V ; Camparo, P

Abstract: Our aims were to analyze reporting of Gleason pattern (GP) 3 and 4 prostate cancer with the ISUP 2005 Gleason grading and to collect consensus cases for standardization. We scanned 25 prostate biopsy cores diagnosed as Gleason score (GS) 6-7. Fifteen genitourinary pathologists graded the digital slides and circled GP 4 and 5 in a slide viewer. Grading difficulty was scored as 1-3. GP 4 components were classified as type 1 (cribriform), 2 (fused), or 3 (poorly formed glands). A GS of 5-6, 7 (3 + 4), 7 (4 + 3), and 8-9 was given in 29%, 41%, 19%, and 10% (mean GS 6.84, range 6.44-7.36). In 15 cases, at least 67% of observers agreed on GS groups (consensus cases). Mean interobserver weighted kappa for GS groups was 0.43. Mean difficulty scores in consensus and non-consensus cases were 1.44 and 1.66 ($p = 0.003$). Pattern 4 types 1, 2, and 3 were seen in 28%, 86%, and 67% of GP 4. All three coexisted in 16% (11% and 23% in consensus and non-consensus cases, $p = 0.03$). Average estimated and calculated %GP 4/5 were 29% and 16%. After individual review, the experts met to analyze diagnostic difficulties. Areas of GP 4 and 5 were displayed as heat maps, which were helpful for identifying contentious areas. A key problem was to agree on minimal criteria for small foci of GP 4. In summary, the detection threshold for GP 4 in NBX needs to be better defined. This set of consensus cases may be useful for standardization.

DOI: https://doi.org/10.1007/s00428-011-1106-x

Interactive Digital Slides With Heat Maps:

A Novel Method to Improve the Reproducibility of Gleason Grading

Lars Egevad, MD, PhD[1], Ferran Algaba, MD[2], Daniel M Berney, MD[3], Liliane Boccon-Gibod, MD[4], Eva Compérat, MD[5], Andrew J Evans, MD[6], Rainer Grobholz, MD, PhD[7], Glen Kristiansen, MD, PhD[8], Cord Langner, MD, PhD[9], Gina Lockwood, PhD[10], Antonio Lopez-Beltran, MD, PhD[11], Rodolfo Montironi, MD, FRCPath[12], Pedro Oliveira, MD[13], Matthias Schwenkglenks, PhD [14], Ben Vainer, MD[15], Murali Varma, MD[16], Vincent Verger[17], Philippe Camparo, MD[18]

[1]Karolinska Institutet, Stockholm, Sweden
[2]Fundacio Puigvert-University Autonomous, Barcelona, Spain
[3]Institute of Cancer, St Bartholomew's Hospital, Queen Mary, University of London, London, United Kingdom
[4]Hopital Armand Trousseau, Paris, France
[5]Hopital La Pitié-Salpetrière, Paris, France
[6]University of Toronto, Toronto, Canada
[7]Kantonsspital Aarau, Aarau, Switzerland
[8]University Hospital, Zurich, Switzerland
[9]Medical University, Graz, Austria
[10]Canadian Partnership Against Cancer, Toronto, Canada
[11]Cordoba University Medical School, Cordoba, Spain
[12]Polytechnic University of the Marche Region, Ancona, Italy
[13]Hospital da Luz, Lisboa, Portugal
[14]University of Basel, Basel, Switzerland
[15]Rigshospitalet, Copenhagen, Denmark
[16]University Hospital of Wales, Cardiff, United Kingdom
[17]CCITI, Dijon, France
[18]Hopital Foch, Paris, France

Address of correspondence:
Lars Egevad, MD, PhD
Dept of Oncology-Pathology, Karolinska Institutet
Radiumhemmet P1:02
Karolinska University Hospital
171 76 Stockholm
Sweden
Phone: +46-8 5177 5492
Fax: +46-8 5177 4524
E-mail: lars.egevad@ki.se

Short title: Digital slides for standardization of prostate cancer grading
Word count: Abstract: 248, Main text: 3024

Abstract

Our aims were to analyze reporting of Gleason pattern (GP) 3 and 4 prostate cancer with the ISUP 2005 Gleason grading and to collect consensus cases for standardization. We scanned 25 prostate biopsy cores diagnosed as Gleason score (GS) 6-7. Fifteen genitourinary pathologists graded the digital slides and circled GP 4 and 5 in a slide viewer. Grading difficulty was scored as 1-3. GP 4 components were classified as Type 1 (cribriform), 2 (fused) or 3 (poorly formed glands). A GS 5-6, 7 (3+4), 7 (4+3), 8-9 was given in 29%, 41%, 19% and 10% (mean GS 6.84, range 6.44-7.36). In 15 cases, at least 67% of observers agreed on GS groups (consensus cases). Mean interobserver weighted kappa for GS groups was 0.43. Mean difficulty scores in consensus and non-consensus cases were 1.44 and 1.66 ($p = 0.003$). Pattern 4 Types 1, 2 and 3 were seen in 28%, 86% and 67% of GP 4. All three co-existed in 16% (11% and 23% in consensus and non-consensus cases, $p = 0.03$). Average estimated and calculated %GP 4/5 were 29% and 16%. After individual review, the experts met to analyze diagnostic difficulties. Areas of GP 4 and 5 were displayed as heat maps, which were helpful for identifying contentious areas. A key problem was to agree on minimal criteria for small foci of GP 4. In summary, the detection threshold for GP 4 in NBX needs to be better defined. This set of consensus cases may be useful for standardization.

Key words: prostate cancer, biopsy, Gleason grading, digital pathology, reproducibility, consensus

Introduction

In 2005 the International Society of Urological Pathology (ISUP) organized a consensus conference on Gleason grading of prostate cancer [1]. Since its inception in the 1960s, the Gleason grading has undergone a gradual transition in its practical application. The purpose of the meeting was to reach a consensus as to how the grading system should be used in needle biopsy (NBX) and radical prostatectomy (RP) specimens. The conference addressed issues in the interpretation of morphological patterns and how to summarize and report grade information.

It is not clear how the ISUP revision of Gleason grading has affected the pathology community. Multiple studies were performed on grading reproducibility in the years prior to the ISUP consensus conference [2-6], but it is unknown how reproducible the modified Gleason grading is among experts in urological pathology. There is also a need to set a standard among general pathologists in application of the new guidelines.

In previous reproducibility studies on prostate cancer grading, the observers assigned Gleason scores, but to our knowledge no attempts were made to analyze which areas the grading decisions were based upon. In this study, 15 experts in urological pathology were asked to circle areas of Gleason patterns (GP) 4 and 5. This enabled an analysis of how grading decisions are made and identification of controversies and agreements that are not reflected by the Gleason scores (GS) alone. The aims were both to assess reproducibility and to better define the critical transition between GP 3 and 4. Interpretation difficulties were analyzed. Our purpose was not only to study areas of disagreement but also to collect consensus cases

for standardization of grading. By publication of such cases, it is possible to minimize the interobserver variation among pathologists.

Materials and Methods

A total of 220 prostate NBX cases were reported at the Karolinska University Hospital, Stockholm, Sweden in January and February 2008, 99 of them as prostate cancer. A set of NBX cores diagnosed as GS 6-7 prostate cancer was selected. Only GS 6 cases that were borderline to GS 7 were included, i.e. cases in which a GS 7 or higher theoretically might be diagnosed. One representative core was chosen from each of 30 biopsy sets, recut and stained with hematoxylin and eosin. The slides were reviewed by one of the authors (L.E.). Cases with technical artifacts were omitted and 27 slides were scanned in a digital slide scanner (CCITI, Dijon, France). Among them two were excluded because of unsatisfactory image quality, leaving 25 digital slides for analysis.

Fifteen experts in urological pathology from 11 countries were invited to participate: Austria (1), Canada (1), Denmark (1), France (3), Germany (1), Italy (1), Portugal (1), Spain (2), Sweden (1), Switzerland (1) and the United Kingdom (2). The experts were asked to assign primary and secondary Gleason grades to the digital slides and a GS was automatically generated. Grading difficulty was scored as 1-3. Score 1 meant that in the expert's opinion only one GS was possible, score 2 that another GS was considered but the pathologist was convinced that the selected GS was the best choice and score 3 meant that the observer was

uncertain if the selected GS was the best score and another score would also be acceptable. GP 4 components were classified as Type 1 (cribriform), 2 (fused) or 3 (poorly formed glands) (**Figure 1A-C).** Each expert circled areas of GP 4 and 5 in the slide viewer and were asked to estimate percentage of GP 4 and 5. These percentages were also calculated from the circled areas. In the image analyzer, the circled areas were layered on top of each other to obtain a summation area, similar to a heat map showing the frequency with which a certain GP was assigned within the biopsy core **(Figure 1D-F).** The darker the color, the more often a GP had been assigned. Green was used for GP 4 and yellow for GP 5. GP 3 was not marked on the digital slides.

After individual review, 11 of the experts met in Paris in October 2009 to analyze diagnostic difficulties and agree on a set of consensus cases (L.B.G., P.C., E.C., D.B., L.E., R.G., G.K., C.L., P.O., B.V., M.V.). The grades were analyzed by individual Gleason scores and by GS categories 5-6, 3+4=7, 4+3=7, 8 and 9.

Statistical analysis

Unpaired Student's t-test was used for comparison of means and chi-2 test for comparison of proportions. A p value less than 0.05 was considered significant. Unweighted and weighted kappa statistics was used to measure interobserver reproducibility.

Results

The original GS distribution was 6 (10, 40%), 3+4=7 (11, 44%), 4+3=7 (4, 16%), respectively (mean 6.6). The 15 experts assigned a GS of 5 (4, 1%), 6 (106, 28%), 3+4=7 (154, 41%), 4+3=7 (72, 19%), 8 (24, 6%) and 9 (15, 4%), respectively (mean 6.84, range 6.44-7.36). The observers reported a GS of 5 in 0-12%, 6 in 0 -52%, 7 in 44-100%, 8 in 0-24% and 9 in 0-16%.

The unweighted kappas of presence of GP 4 and 5 were 0.54 and 0.81, respectively. The weighted kappas of Gleason scores 5-9 and of the GS categories 5-6, 3+4=7, 4+3=7, 8 and 9 were 0.35 and 0.43, respectively.

An average of 74% of the experts agreed on the most commonly assigned GS (5, 6, 7, 8 or 9) with a range of 47-93% in the 25 cases. A 100% concordance was not reached for any of the cases either for the Gleason scores or the GS categories. A concordance of at least 80% was obtained in 12 cases for Gleason scores and in 6 cases for the GS categories. A concordance of at least 67% was obtained in 19 cases for Gleason scores. In 15 cases there was a concordance of at least 67% for the GS categories 5-6, 3+4=7, 4+3=7, 8 and 9 and these were designated as consensus cases. Their Gleason scores were 6, 3+4=7 and 4+3=7 in 6, 7 and 2 cases, respectively **(Table 1)**. A concordance of 33-60% was reached for GS categories in 10 cases, which were designated as non-consensus cases **(Table 2)**. In one of these cases the spread of GS was from 6 to 9 and in two other cases from 6 to 8.

Difficulty scores were reported by 13 of the 15 observers. A difficulty score of 1, 2 or 3 was assigned in 58%, 32% and 10%, respectively **(Table 2)**. The average reported difficulty score

was 1.53 (range 1.20-1.88 among 13 observers). Eight observers had a median difficulty score of 1 and 5 had a median score of 2. In the 15 consensus cases the average difficulty score was 1.44 as compared to 1.66 in the 10 non-consensus cases (p = 0.003). Among consensus cases a difficulty score of 1, 2 or 3 was assigned in 65%, 27% and 8%, respectively. Among non-consensus cases a difficulty score of 1, 2 or 3 was assigned in 48%, 38% and 14%, respectively. When biopsies were assigned a difficulty score 3, 53% (18 of 34) were non-consensus cases compared to 33% (62 of 188) when a score 1 was assigned (p = 0.026).

Thirteen observers categorized the GP 4 according to type. GP 4 Type 1 (cribriform), 2 (fused) or 3 (poorly formed glands) were seen in 20%, 61% and 47%, respectively **(Table 4)**. A GP 4 was reported in 70% (228 of 325). When a GP 4 was reported, Types 1, 2 and 3 were seen in 28%, 86% and 67%. When a GP 4 was reported, all three patterns, two patterns and one pattern were seen in 16%, 50% and 34%, respectively **(Table 5)**. All three patterns were seen in 11% of the consensus cases and 23 % of the non-consensus cases (p = 0.03). Per case, 1.88, 6.88 and 2.56 observers assigned a GP type of 1, 2 and 3, respectively. A GP 5 was assigned by at least one observer in seven cases (by one to eight observers).

Regions were marked by 14 of 15 observers. The precision of marking the areas was highly variable. The mean percentage of GP 4 cancer was 28% when estimated subjectively and 16% when calculated by the image analyzer. The mean percentage of GP 5 was 0.5% as estimated and 0.1% as calculated.

Areas of GP 4 and 5 were displayed as heat maps with overlaying regions. The maps were helpful for identifying contentious areas. A key problem was to agree on minimal criteria for small foci of GP 4 **(Figure 2).** In an open discussion, areas were identified where the majority

had overlooked a minute component of cancer that might be interpreted as high-grade carcinoma. These included occasional solid strands **(Figure 2A-B)** or fusion patterns **(Figure 2C-D)** in GS 3 + 3 = 6, which could be considered tangential sections of GP 3 cancer glands. In other cases, such structures were too abundant to be overlooked **(Figure 2E-F)**, leading to a consensus grading of GS 3 + 4 = 7.

Discussion

Over the past decades there has been a gradual change of the practice of Gleason grading, causing considerable grade shift upwards [7]. The recent ISUP revision of the Gleason grading system has contributed to this upgrading. Helpap *et al*. showed that the number of NBX cases diagnosed as GS 6 decreased from 48% to 22% when adopting the ISUP modification of Gleason grading, while the number of GS 7 cases increased from 26% to 68% [8]. In a recent study by Delahunt *et al*. on patients with locally advanced prostate cancer, the number of cases diagnosed as GS 6 decreased from 11% to 5% when going from conventional to ISUP modified GS [9].

Reasons for upgrading include both changed pattern interpretation and new recommendations for how to report the GS on NBX specimens. The ISUP guidelines proposed that GP 1 should be avoided 'with extremely rare exception', that GS 2+2= 4 should rarely be diagnosed on NBX and that most cribriform cancers should be diagnosed as GP 4 rather than 3 [1]. Furthermore, cancers with incomplete, poorly formed glands were now included in GP 4. Some of these rules for pattern interpretation were implemented by many experts already before the ISUP revision. Perhaps more important, there was previously a widespread

tradition to overlook small foci of higher grade or to diagnose them as tertiary pattern of higher grade, but not include them in the GS. Prior to the ISUP consensus meeting a GP of higher grade was generally included only if more than 5% of all cancer present [10]. One of the key recommendations of the ISUP meeting was that even minute foci of higher grade should be included in the GS on NBX specimens, regardless of the extent though it was added they should be identified at 'low to medium power'. This possible equivocation has probably meant that, pathologists may have different detection thresholds for small high-grade components and there is a need to standardize this interpretation. In retrospect, this part of the ISUP revision of Gleason grading can be questioned. The consequences have evidently been a substantial upgrading in some institutions. There is now considerable uncertainty among pathologists where to place the threshold for recognition of minute high-grade components. There is also a concern among urologists how a GS 7 cancer on needle biopsy should now be treated. It would have been very helpful if the ISUP consensus meeting 2005 had presented more precise practical guidelines regarding GP 4 detection with e.g. ample illustrations. The best that can be done in the current situation is that experts attempt to agree on a standardized interpretation of the GP 3 to 4 transition. As a first step we aim to use the microphotograph material from this study to set up a web-based image library that is made accessible for European pathologists.

The weighted kappa of Gleason scores in the present study was lower than reported in many previous studies on interobserver reproducibility of Gleason grading [2, 3, 5, 6]. However, the results are not comparable as the biopsy cores of the present study were selected to include borderline cases between GS 6 and 7. The unweighted kappas of presence of GP 4 and particularly of GP 5 were far better. The latter may be an over estimation as the cases were primarily selected to lack a GP 5 component. There are indeed multiple pitfalls with the use of

kappa statistics. How study cases are selected is of critical importance. The safest method to avoid selection bias is to use a consecutive series. Unfortunately, it is not feasible to make such a study on a consecutive series that is sufficiently large to include difficult and interesting cases. More commonly, there is an accumulation of cases that experts consider unequivocal so the golden standard of the study cannot be criticized. By contrast, in the present study, we have selected cases with borderline morphology. The easiest cases of GS 6 were excluded and only GS 6 cases where it was thought that someone might diagnose a GS 7 were included. We have also refrained from inclusion of classical GS 9-10 cases, which would also most likely give a high grading reproducibility. Thus, it is not surprising that the kappa statistics in our study were comparatively low and they should rather be used for comparisons within the study.

We here used digitized slides that were read in a slide viewer. It has been shown that the diagnostic accuracy using digital slides is similar to conventional glass slides [11]. It would not have been possible to circulate glass slides to 15 experts in 11 countries within a reasonable time frame. The other advantage with digital slide is that they allow marking of regions of interest. To our knowledge this is the first attempt to identify the areas that Gleason grading decisions are based upon. The advantage of the heat map system is that it enables an analysis of which patterns are controversial and which are easier to grade. After an independent assessment of grades a majority of the group met in Paris to discuss controversies and analyze why there was disagreement in some cases. In the discussion the heat maps could be removed from the digital slides to show the detailed histological morphology. While doing this we compared morphological features with lists of individual grading results. It was sometimes evident that disagreement was caused by an occasional outlier, while in other cases, the participants were split into more equal groups of different standpoints. Certain

patterns seemed to be more difficult than others, such as incomplete glands where the detection threshold was critical for accurate grading.

We found that there is not only a variability in how Gleason scores are assigned but also in terms of what precise areas decisions are based on. When minute foci of higher grade should be included in the GS on NBX specimens, regardless of the extent, it is necessary to define a detection threshold in order to avoid the degeneration of Gleason grading into a three-tiered grading system running from GS 7 to 9. In the open discussion, such examples were found to include possible tangential cuts of complete or incomplete glandular structures.

Fusion pattern was found to be the most common GP 4 type, while cribriform cancer, i.e. one of the classical variants of GP 4, was least common. A novel feature of the ISUP revision is that poorly formed or incomplete glands are included in GP 4, which is in line with the general concept of GP 4 as cancer that attempts to form glands, yet falls short of forming complete, circumscribed glands. Poorly formed glands were seen in 47% of GP 4, making it the next most common type of GP 4 after the fusion pattern. All three patterns were seen in 16% of GP 4 cases, more commonly in non-consensus cases than in consensus cases, indicating that the presence of a heterogeneous GP 4 component adds to the grading difficulty.

The purpose of the study was primarily to define the transition between Gleason patterns 3 and 4 and the cases were selected accordingly. Yet a GP 5 was assigned by at least 1 observer in 7 cases (by 1 to 8 observers). In a study by Egevad *et al*., a questionnaire was distributed to 91 genitourinary pathologists in countries around the world [12]. Rare individual cells, strands, or nests identified only at less than 40x lens magnification were considered sufficient to diagnose GP 5 on needle biopsy by 17% of pathologists, whereas 83% required clusters of

such structures seen at lower than 40x magnification. The GP 5 identified by occasional observers in some cases proved to be minute components rather than cohesive areas and should be overlooked according to the vast majority of observers.

Interestingly, cases with higher self-rated grading difficulty had lower reproducibility. A difficulty score of 1, 2 or 3 was assigned in 58%, 32% and 10%, respectively, which means that as many as 42% of assessments were accompanied with a certain degree of doubt as to the correct grade. It is on the other hand surprising that in 42% of non-consensus cases, the experts gave a difficulty score of 1, i.e. they were absolutely certain that their grade was the correct one. Thus, there is a discrepancy between our self-confidence and the performance as a group. These results have to be interpreted with some caution as a research study differs from routine pathology work. A limited number of cases are reviewed and the participants are usually devoted to the task. A certain fear to deviate from the midstream of the group may add to the motivation. Thus, study participants are more likely to report a high self-rated grading difficulty in cases where they fail to reach consensus, while in a busy routine practice, mistakes may also be done in cases that are felt to be easy.

Percent high-grade cancer (%GP 4/5) has been suggested as a prognostic factor for prostate cancer [13-16]. The Gleason grading system was originally a nine-tiered system, but only some of the grades are actually commonly used. Therefore, %GP 4/5 may help to separate cases in the mid range of the Gleason scores. A ‘sliding scale’ grading system, similar to nomograms has attractions over a stochastic grading system, which can fall prey to the problems exemplified by this study. The present study is to our knowledge the first time a comparison has been done between subjective and objective assessment of %GP 4/5.

Interestingly, the subjectively estimated area was almost twice as large as the objective measurement. This may be explained by some inherent difficulties in the assessment of proportions of GPs. Prostate cancer specimens contain a mixture of malignant and benign glands and also of epithelial and stromal components. When there are multiple GPs present, eg. GP 3 and GP 4, they are often mixed. The assessment of areas occupied by a certain GP is thus not quite straight-forward. It may vary between individuals how much of non-cancerous elements within the tumor we include or subtract. The higher percentages by subjective assessment may indicate that the human eye tends to overlook presence of eg. occasional GP 3 glands within a GP 4 or the presence of benign glands within the tumor. When we are asked to document our assessment by outlining a high-grade component digitally, it seems to make us more aware of such non-high-grade components. On the other hand, the subjective estimation of %GP 4/5 has been shown to be at least as reproducible as that of the GS. In an interobserver reproducibility study, 4 observers had a mean weighted kappa for biopsy GS and %GP 4/5 of 0.48 to 0.55 (overall mean 0.51) and 0.52 to 0.68 (overall mean 0.60), respectively [17]. However, it was found that we subjectively over estimate %GP 4/5 as compared to computerized calculation of the percentage. A difficulty in the assessment of percentages of Gleason patterns is that malignant glands are often mixed with stroma and benign glands and it is difficult to know how much of intervening tissues that should be subtracted.

The clinical performance of GS as predictor of outcome is evidently most important when determining the optimal definition of grade. Billis *et al*. showed that both conventional and modified GS on NBX predicted outcome after RP [18]. A lower p value was obtained with the modified GS and the authors concluded that the conventional grading was outperformed. However, the Kaplan Meier curves for GS 6 and 7 were almost identical with modified

grading and the separation seems to be between GS 7 vs. 8 or higher. Delahunt *et al*. recently questioned the performance of the modified GS as predictor of disease progression [9]. Further studies are needed and it is important that those studies clearly define how the GS was obtained. A major disadvantage of the ISUP 2005 revision of the Gleason grading system was that it did not include any validation studies comparing the revised and the conventional versions of this grading system.

Perhaps the most useful part of this study was the identification of a set of 15 cases in which there was a 67% consensus among experts in terms of GS categories. This set of cases may serve as a library for standardization among general pathologists. For this purpose, we intend to use the European Network of Uropathology (ENUP), an international network of communication recently organized by the Uropathology Working Group of the European Society of Pathology (ESP) [19]. The purpose of ENUP is to establish a channel for distribution of information about urological pathology such as guidelines, consensus documents, meetings and courses, to organize research collaborations and to set up mechanisms for survey studies. ENUP has recruited a total of 374 individual members from 338 pathology laboratories in 15 west European countries. Email is used for all communication and studies are carried out through interactive websites. Our intention is to invite the ENUP members to use the same set of cases as in the expert study for a larger reproducibility study and then give feedback as to how experts agreed on the same cases. Thereby, it will be possible to use a library of digital slides for stadardization of histopathological grading.

In summary, there is still a considerable disagreement among experts on borderline GS 6-7 cases, using the ISUP revision of the Gleason grading of prostate cancer. A major source of

disagreement is that the detection threshold for minimal foci of GP 4 in NBX needs to be better defined. Yet, we were able to identify a set of consensus cases that are likely to be useful for standardization of Gleason grading.

Acknowledgment

D.M.B. is supported by Orchid.

Legends

Table 1. Fifteen consensus cases with at least 67% agreement for GS categories. Number of votes for most commonly assigned grade shown in bold font.

Table 2. There were 10 non-consensus cases with agreement for GS categories in less than 67%. Number of votes for most commonly assigned grade shown in bold font.

Table 3. Difficulty scores in 15 consensus and 10 non-consensus cases. Score 1 = only one Gleason score (GS) was possible, score 2 = another GS was considered but the pathologist was convinced that the selected GS was the best choice and score 3 = uncertain if the selected GS was the best score.

Table 4a. Gleason pattern (GP) 4 types in consensus and non-consensus cases. GP4 Type 1 = cribriform glands, type 2 = fused glands and type 3 = poorly formed glands. Percentages of all cases.

Table 4b. Gleason pattern (GP) 4 types in consensus and non-consensus cases. GP4 Type 1 = cribriform glands, type 2 = fused glands and type 3 = poorly formed glands. Percentages of cases with a GP 4 reported.

Table 5. Number of Gleason pattern (GP) 4 types in consensus and non-consensus cases. Percentages of cases with a GP 4 reported.

Figure 1A. Gleason pattern (GP) 4 of cribriform type (Type 1). **B.** GP 4 of fusion type (Type 2). **C.** GP 4 with incomplete glands (Type 3). 20X lens magnification in A - C. **D.** Needle biopsy with cancer encircled by red line. Overview to the left, medium slide viewer magnification to the right. **E.** Areas of GP 4 marked with green lines by individual observers. **F.** Marked areas were added on top of each other to produce a heat map showing the frequency with which a certain GP was assigned within the biopsy core. The darker green, the more often a GP 4 was assigned.

Figure 2A. Consensus case of Gleason score (GS) 3+3=6. 10X lens magnification. **B.** The same consensus case of GS 3+3=6. Most glands are circumscribed and well-defined. Arrow indicates a few epithelial strands that can be overlooked, as they possibly represent tangential cuts. To be ignored such structures must be only occasional finding. 20X lens magnification. **C.** Consensus case of GS 3+3=6. 10X lens magnification. **D.** The same consensus case of GS 3+3=6. Arrow indicates a few seemingly fused glands. When only occasional merged glandular structures are found, they should not be interpreted as a fusion pattern, as they possibly represent tangential cuts. 20X lens magnification. **E.** Consensus case of GS 3+4=7. 10X lens magnification. **F.** The same consensus case of GS 3+4=7. Some glands are circumscribed and well-defined. However, arrow indicates poorly formed glands that are too abundant to be and dispersed throughout the visual field to be overlooked. The 2005 ISUP consensus meeting decided to include such poorly formed or incomplete glands in GP 4. 20X lens magnification.

Acknowledgement

Sources of support: D.M.B. is supported by Orchid.

Conflict of interest statement

We declare that we have no conflict of interest.

References

1. Epstein JI, Allsbrook WC, Jr., Amin MB, Egevad LL (2005) The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma Am J Surg Pathol 29:1228-1242

2. Allsbrook WC, Jr., Mangold KA, Johnson MH, Lane RB, Lane CG, Amin MB, Bostwick DG, Humphrey PA, Jones EC, Reuter VE, Sakr W, Sesterhenn IA, Troncoso P, Wheeler TM, Epstein JI (2001) Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists Hum Pathol 32:74-80.

3. Allsbrook WC, Jr., Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI (2001) Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist Hum Pathol 32:81-88.

4. Carlson GD, Calvanese CB, Kahane H, Epstein JI (1998) Accuracy of biopsy Gleason scores from a large uropathology laboratory: use of a diagnostic protocol to minimize observer variability Urology 51:525-529

5. Egevad L (2001) Reproducibility of Gleason grading of prostate cancer can be improved by the use of reference images Urology 57:291-295.

6. Lessells AM, Burnett RA, Howatson SR, Lang S, Lee FD, McLaren KM, Nairn ER, Ogston SA, Robertson AJ, Simpson JG, Smith GD, Tavadia HB, Walker F (1997) Observer variability in the histopathological reporting of needle biopsy specimens of the prostate Hum Pathol 28:646-649

7. Berney DM, Fisher G, Kattan MW, Oliver RT, Moller H, Fearn P, Eastham J, Scardino P, Cuzick J, Reuter VE, Foster CS (2007) Major shifts in the treatment and prognosis of prostate cancer due to changes in pathological diagnosis and grading BJU Int 100:1240-1244

8. Helpap B, Egevad L (2006) The significance of modified Gleason grading of prostatic carcinoma in biopsy and radical prostatectomy specimens Virchows Arch 449:622-627

9. Delahunt B, Lamb DS, Srigley JR, Murray JD, Wilcox C, Samaratunga H, Atkinson C, Spry NA, Joseph D, Denham JW (2010) Gleason scoring: a comparison of classical and modified (international society of urological pathology) criteria using nadir PSA as a clinical end point Pathology 42:339-343

10. Deshmukh N, Foster CS (1998) Grading prostate cancer. In: Foster CS, Bostwick DG (eds) Pathology of the Prostate. WB Saunders, Philadelphia, pp. 191-227

11. Koch LH, Lampros JN, Delong LK, Chen SC, Woosley JT, Hood AF (2009) Randomized comparison of virtual microscopy and traditional glass microscopy in

diagnostic accuracy among dermatology and pathology residents Hum Pathol 40:662-667

12. Egevad L, Allsbrook WC, Jr., Epstein JI (2005) Current practice of Gleason grading among genitourinary pathologists Hum Pathol 36:5-9

13. Egevad L, Granfors T, Karlberg L, Bergh A, Stattin P (2002) Percent Gleason grade 4/5 as prognostic factor in prostate cancer diagnosed at transurethral resection J Urol 168:509-513.

14. Stamey TA, McNeal JE, Yemoto CM, Sigal BM, Johnstone IM (1999) Biological determinants of cancer progression in men with prostate cancer JAMA 281:1395-1400

15. Stamey TA, Yemoto CM, McNeal JE, Sigal BM, Johnstone IM (2000) Prostate cancer is highly predictable: a prognostic equation based on all morphological variables in radical prostatectomy specimens J Urol 163:1155-1160

16. Vis AN, Roemeling S, Kranse R, Schroder FH, van der Kwast TH (2007) Should we replace the Gleason score with the amount of high-grade prostate cancer? Eur Urol 51:931-939

17. Glaessgen A, Hamberg H, Pihl CG, Sundelin B, Nilsson B, Egevad L (2003) Interobserver reproducibility of percent Gleason grade 4/5 in prostate biopsies J Urol 171:664-667

18. Billis A, Guimaraes MS, Freitas LL, Meirelles L, Magna LA, Ferreira U (2008) The impact of the 2005 international society of urological pathology consensus conference on standard Gleason grading of prostatic carcinoma in needle biopsies J Urol 180:548-552; discussion 552-543

19. Egevad L, Algaba F, Berney DM, Boccon-Gibod L, Griffiths DF, Lopez-Beltran A, Mikuz G, Varma M, Montironi R (2009) The European Network of Uropathology: a novel mechanism for communication between pathologists Anal Quant Cytol Histol 31:90-95

Tables

	Gleason score categories					
Case	**5-6**	**7 (3+4)**	**7 (4+3)**	**8**	**9**	**Agreement (%)**
1	**12**	3	0	0	0	80
2	0	**10**	2	3	0	67
4	2	**10**	3	0	0	67
6	2	**11**	1	0	1	73
10	**13**	2	0	0	0	87
12	1	**11**	1	2	0	73
14	0	1	**13**	0	1	87
15	1	**10**	3	1	0	67
16	0	**13**	1	1	0	87
17	**11**	3	1	0	0	73
18	**10**	4	1	0	0	67
19	**11**	4	0	0	0	73
21	1	0	**11**	3	0	73
22	1	**14**	0	0	0	93
23	**12**	3	0	0	0	80

Table 1. Fifteen consensus cases with at least 67% agreement for Gleason score (GS) categories. Number of votes for most commonly assigned grade shown in bold font.

	Gleason score categories					
Case	**5-6**	**7 (3+4)**	**7 (4+3)**	**8**	**9**	**Agreement (%)**
3	**8**	7	0	0	0	53
5	2	**5**	3	2	3	33
7	0	6	**8**	1	0	53
8	0	6	**8**	1	0	53
9	2	**9**	1	3	0	60
11	0	2	**8**	3	2	53
13	**8**	7	0	0	0	53
20	6	**7**	2	0	0	47
24	0	0	4	3	**8**	53
25	**7**	6	1	1	0	47

Table 2. Ten non-consensus cases with agreement for Gleason score (GS) categories in less than 67%. Number of votes for most commonly assigned grade shown in bold font.

	Difficulty scores					
	1	**2**	**3**	**Total**	**Mean**	**p value**
Consensus	65% (126)	27% (53)	8% (16)	195	1.44	0.003
Non-consensus	48% (62)	38% (50)	14% (18)	130	1.66	
Total	58% (188)	32% (103)	10% (34)	325		
p value						

Table 3. Difficulty scores in 15 consensus and 10 non-consensus cases. Score 1 = only one Gleason score (GS) was possible, score 2 = another GS was considered but the pathologist was convinced that the selected GS was the best choice and score 3 = uncertain if the selected GS was the best score.

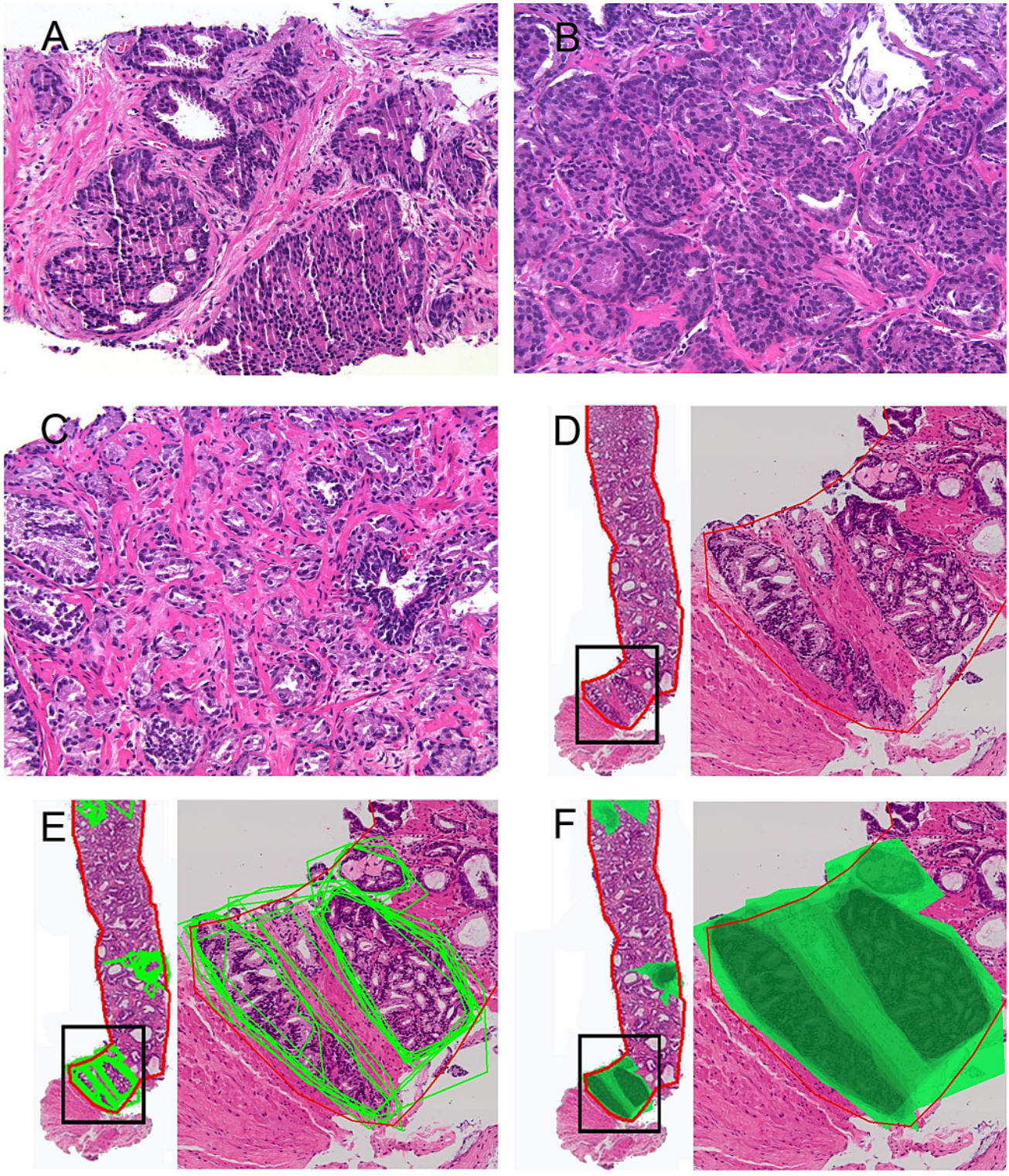
	Type of Gleason pattern 4				
	1	**2**	**3**	**Total**	**p value**
Consensus	24% (31)	81% (106)	67% (88)	131	0.38
Non-consensus	35% (34)	94% (91)	66% (64)	97	
Total	28% (65)	86% (197)	67% (152)	228	

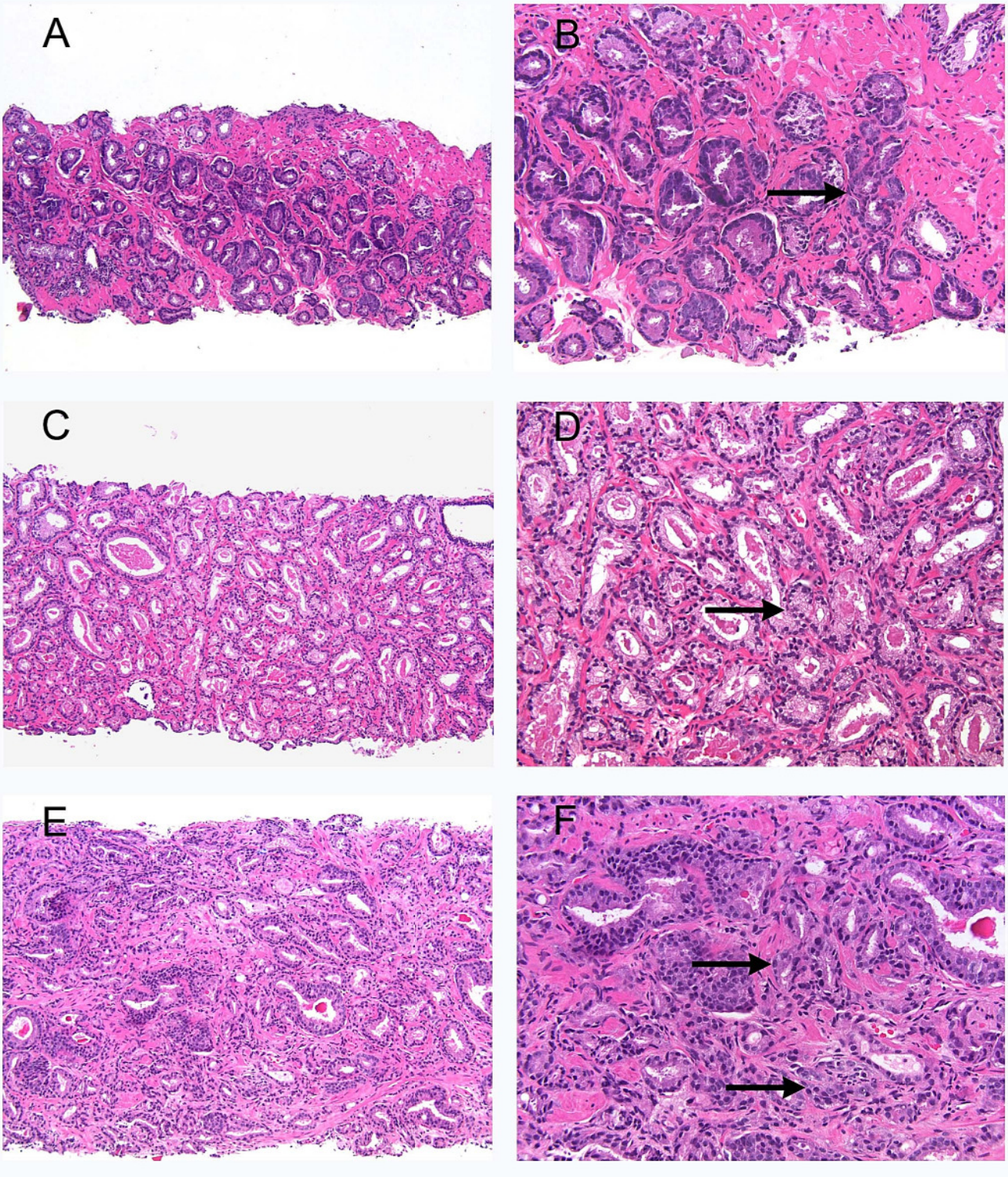
Table 4. Gleason pattern (GP) 4 types in consensus and non-consensus cases. GP4 Type 1 = cribriform glands, type 2 = fused glands and type 3 = poorly formed glands. Percentages of cases with a GP 4 reported.

	Number of Gleason Pattern 4 Types				
	1	**2**	**3**	**Total**	**p value**
Consensus	39% (51)	50% (66)	11% (14)	131	0.03
Non-consensus	28% (27)	50% (48)	23% (22)	97	
Total	34% (78)	50% (114)	16% (36)	228	

Table 5. Number of Gleason pattern (GP) 4 types in consensus and non-consensus cases.

Percentages of cases with a GP 4 reported.


A
B
C
D
E
F


A
B
C
D
E
F